A chromosome-based model for estimating the number of conserved segments between pairs of species from comparative genetic maps

A Chromosome-Based Model for Estimating the Number of Conserved Segments Between Pairs of Species From Comparative Genetic Maps

David Waddington, Anthea J. Springbett and David W. Burt

Roslin Institute (Edinburgh), Roslin, Midlothian EH25 9PS, Scotland

Manuscript received April 23, 1999
Accepted for publication September 20, 1999

ABSTRACT

Comparative genetic maps of two species allow insights into the rearrangements of their genomes since divergence from a common ancestor. When the map details the positions of genes (or any set of orthologous DNA sequences) on chromosomes, syntenic blocks of one or more genes may be identified and used, with appropriate models, to estimate the number of chromosomal segments with conserved content conserved between species. We propose a model for the distribution of the lengths of unobserved segments on each chromosome that allows for widely differing chromosome lengths. The model uses as data either the counts of genes in a syntenic block or the distance between extreme members of a block, or both. The parameters of the proposed segment length distribution, estimated by maximum likelihood, give predictions of the number of conserved segments per chromosome. The model is applied to data from two comparative maps for the chicken, one with human and one with mouse.

COMPARATIVE gene mapping, the analysis of the chromosomal location of homologous genes in different species, is a powerful tool for gene mapping and the study of genome organization and evolution. The most detailed comparisons are between mouse and man, with >2000 homologous genes mapped in both species. Almost 200 linkage groups are conserved between these two species (Carver and Stubbs 1997). Even before these detailed comparative gene maps were assembled, the early genetic maps of man and mouse were used to estimate the mean length and number of chromosomal segments conserved during evolution (Nadeau and Taylor 1984). Comparison of the locations of 83 homologous loci revealed 13 conserved segments. Statistical models were developed for using this sample of conserved segments to estimate the mean length of all conserved autosomal segments in the genome as 8.1 cM. This was used to estimate the number of conserved segments as 198, which is very close to the number observed today. Most comparative studies have focused on mammals, notably mouse and human comparisons (O'Brien *et al.* 1993, 1997; Womack and Kata 1995; Andersson *et al.* 1996; Carver and Stubbs 1997). Recently, comparisons between birds (Burt *et al.* 1995; Andersson *et al.* 1996; Jones *et al.* 1997; Pitel *et al.* 1998; Smith and Cheng 1998) or bony fish (Morizot 1983; Postlethwait *et al.* 1998) and mammals reveal a high degree of conservation of genome organization. This is surprising given that these species diverged from a common ancestor 420 mya.

Corresponding author: Dave Waddington, Roslin Institute (Edinburgh), Roslin, Midlothian EH25 9PS, Scotland.
E-mail: dave.waddington@bbsrc.ac.uk

The genetic marker maps of farm animals such as cattle, pigs, and poultry are now sufficiently well advanced to be of practical value for the study of economically important traits and livestock improvement (Andersson *et al.* 1996). Knowledge of the location of coding sequences is, however, limited. Maps of major livestock species contain 1000–2000 anonymous microsatellite markers and only 5–10% of all genetic markers are genes. Mapping of several vertebrate genomes is progressing rapidly, but by far the most detailed information is still to be found for mouse and human. Through comparative gene mapping, it is possible to link the "gene-poor" maps of livestock to the "gene-rich" maps of human and mouse (Andersson *et al.* 1996).

Many measures of genome rearrangement are possible, depending on the level of gene mapping information available (*e.g.*, synteny, gene order, and gene position) and the corresponding mathematical modeling approach used. Two derived measures of the degree of genome reorganization between two species using synteny data have been proposed (Bengtsson *et al.* 1993), and also a measure of genome similarity using gene order (Zakharov *et al.* 1995). More mechanistic models have been derived from some or all of the known chromosome modification mechanisms such as reciprocal translocation, inversion, transposition, and chromosome fusion and fission. Such an approach has been developed to obtain a direct estimate of the number of conserved segments from synteny data (Sankoff and Nadeau 1996; Erlich *et al.* 1997), which takes account of as yet unobserved syntenies. When sequences of genes are accurately mapped, similar descriptive models of genome rearrangement are possible (Sankoff 1993; Hannenhalli 1995; Hannenhalli and Pevzner 1995)

but these models do not allow for undiscovered segments.

Our concern is with incomplete data of an intermediate accuracy arising from genetic maps, which yield blocks of conserved synteny. These contain information on the number of genes per block and the measured distance (or range) between extreme genes in blocks with at least two members, but ignore information on gene order. The first published estimate of the number of conserved segments between man and mouse used an approach based on such data (Nadeau and Taylor 1984). Although this landmark work used measurements of distance, subsequent approaches have concentrated either on counts of genes (by chromosome or blocks within chromosomes) or on gene order. We build on the approach of Nadeau and Taylor (1984) by using both counts and the additional distance information available in ranges, when present. A central assumption of the Nadeau and Taylor (1984) model was that all chromosomes had identical distributions of the lengths of segments from which ranges had been sampled. Chromosome lengths were assumed to be large relative to segment lengths. This approximation is good for chromosomes >100 cM in length, and fair for those >50 cM in length (as in the mouse), but is untenable for species with shorter chromosomes, such as the chicken, which has extreme divergence in chromosome size. The currently established chicken linkage group sizes range from 2 cM to 518 cM, with several <50 cM. We have extended the method of Nadeau and Taylor (1984) to allow small chromosome lengths and also to use the probability density of the observed ranges in a likelihood approach. A similar method, using only the number of genes forming a syntenic block of one or more markers, is also proposed. This leads naturally to a combined approach using both types of data. The model allows a flexible description of chromosome breakage, which includes random breakage as a special case. The methods are illustrated using comparative maps that compare chickens with both humans and mice (Burt *et al.* 1999).

METHODS

Distributions of segment lengths for different chromosomes: How are the lengths of conserved segments expected to change, in general, as chromosome length increases? Very small (hypothetical) chromosomes are likely to contain only a single conserved segment, while large chromosomes might be expected to contain many relatively short segments and a few long ones. Intermediate length chromosomes may have segments whose lengths are a substantial proportion of chromosome length. Thus, for our empirical model of segment lengths, we require a flexible distribution whose shape can be defined for each chromosome. The β distribution, a two-parameter distribution defined on the unit interval, can give distributional shapes as varied as unimodal, uniform, exponential-like, and reverse exponential, as its parameters vary. Segment lengths for the kth chromosome can be scaled by chromosome length, l_k, to follow a β distribution whose parameters are a function of l_k. Using the square parenthesis notation for a density function, assume the distribution of segment lengths, y, on chromosome k to be

$$[y]_k = \frac{1}{l_k}\left(\frac{y}{l_k}\right)^{a-1}\left(1 - \frac{y}{l_k}\right)^{b-1} / B(a, b),$$

where a and b are the β distribution parameters and $B(a, b)$ is the β function $\int_0^1 x^{a-1}(1 - x)^{b-1}\,dx$. Assume also that the β parameters change with chromosome length in a smooth way: $a = \alpha l_k^{\beta}$ and $b = \gamma l_k^{\delta}$. The mean segment length on chromosome k is then $l_k a/(a + b)$ and the expected number of segments S_k is $(a + b)/a$, or $S_k = 1 + (\gamma/\alpha) l_k^{(\delta-\beta)}$. The expected number of conserved segments is the sum of S_k over all chromosomes.

An important special case occurs when the β distribution parameter a equals one, so that $[y]_k = b(1 - y/l_k)^{b-1}/l_k$. This is the distribution of segment lengths when there are b random breaks in a chromosome (Sankoff and Nadeau 1996), and thus there are $S_k = b + 1$ conserved segments. Strictly, this random breakage pattern results from the superposition of chromosome breakage patterns of two species. We use the terms *random genome breakage* model and *random chromosome breakage* model to distinguish the following two cases where $\delta = 1$ and $\delta \neq 1$. When $a = 1$ and $\delta = 1$, b is a linear function of l_k. This corresponds to the random breakage model commonly found in the literature, which assumes that chromosome breakage occurs entirely at random throughout the genome with density γ breaks per centimorgan. When $\delta \neq 1$, then the density of random breaks changes from chromosome to chromosome. The more general nonrandom breakage model presented here uses only three parameters. Adding a fourth parameter produces no appreciable improvement in fit to our data. We set $\beta = 0$, and estimate the constants α, γ, and δ. Comparisons of likelihoods from these three models, starting from the three-parameter model and simplifying, allow the plausibility of random breakage models to be assessed.

Count data: Observed genes are assumed to be distributed at random along the genome with constant density D genes per centimorgan. If there are many genes and a large number of observed syntenic groups, then the distribution of the number of genes (n) in a syntenic group found on an underlying conserved segment of length y will be approximately Poisson with mean Dy, defined for observable values of $n \geq 1$.

The distribution of n, given y, is

$$[n|y]_1 = \frac{(Dy)^n \exp(-Dy)}{n!\{1 - \exp(-Dy)\}} \quad \text{for } n \geq 1.$$

The marginal distribution of n is then

$$[n]_k = \int_0^{l_k} [n|y]_1 [y]_k dy.$$

Parameters α, β, γ, and δ (if not fixed) are estimated by maximizing the log-likelihood $L_1 = \Sigma_{\text{all counts}} \log([n]_k)$, with the integral evaluated numerically.

Range data and combined data: An extension of the scheme for counts follows naturally for syntenic groups of at least two genes, where we have additional information on the range, *w*, between the outermost pair of the group. For this subset of the data the Poisson distribution of *n* given *y* has to be truncated to be ≥2:

$$[n|y]_2 = \frac{(Dy)^n \exp(-Dy)}{n!\,\{1 - (1 + Dy)\exp(-Dy)\}} \quad \text{for } n \geq 2.$$

The distribution of the range, conditional on *n* and *y*, is

$$[w|n, y] = \frac{n(n-1)w^{n-2}}{y^{n-1}}\left(1 - \frac{w}{y}\right) \quad \text{for } 0 < w \leq y$$

$$= 0 \qquad \text{for } w < 0 \text{ or } w > y$$

(Plackett 1971). Then the joint distribution of *n* and *w* is

$$[n, w]_k = \int_w^{l_k} [w|n, y][n|y]_2 [y]_k dy.$$

Note that the lower limit of this integral is no longer zero because the underlying segment must be at least as long as the observed range. Parameters are estimated by maximizing the log-likelihood $L_2 = \Sigma_{\text{ranges}} \log([n, w]_k)$.

It is possible to combine both preceding likelihoods. For single loci ($n = 1$) the distribution of *n* in the *Count data* section may be used. For range data the joint distribution of *n* and *w* may be used, with one modification. The Poisson distribution for a count conditional on segment length, $[n|y]_2$, should be truncated to allow $n \geq 1$ rather than $n \geq 2$. This gives a common truncated Poisson distribution for both approaches, so that their respective log-likelihoods may be added.

Then we maximize $L_3 = \Sigma_{\text{single loci}} \log([n]_k) + \Sigma_{\text{ranges}} \log([n, w]_k)$.

Confidence intervals: All maximizations were performed using standard derivative-free optimization routines. Confidence intervals for the number of conserved segments, *S*, were calculated only for the two random breakage models with β parameter $a = 1$.

The random chromosome breakage model has a confidence region for δ and log(γ) that is an elliptical area defined by the critical log-likelihood contour corresponding to $L_{\max} - \chi_2^2(0.95)/2$. For *k* indexing all $N = 38$ autosomes,

$$S = N + \sum_k \gamma l_k^{\delta}$$

and

$$\log(S - N) = \log(\gamma) + \log\left(\sum_k l_k^{\delta}\right).$$

When *S* is fixed at a value S_0, γ and the log-likelihood may be expressed as a function of δ. For small changes in δ the contours of constant *S* are almost linear and run approximately parallel to the major axis of the elliptical confidence region for log(γ) and δ. The likelihood was maximized for δ, over a grid of integer S_0 values, and if the maximum exceeded that of the critical contour, then S_0 was taken to be inside the confidence interval for *S*.

The confidence interval for the random genome breakage model was found from the log-likelihood corresponding to a grid of integer S_0 values, using the critical value $L_{\max} - \chi_1^2(0.95)/2$.

Comparing model and data: Observed genes are assumed to be distributed at random over the genome. Those found by means that are not random (previously mapped by FISH, gene families, chromosome walking, cross-referenced genes from other species' maps, etc.) have been omitted. If the distribution of genes is random and of constant density *D*, then, on average, the number found on linkage group *k* will be proportional to the length of the linkage group, l_k, and the observed number, m_k, will follow a Poisson distribution with mean Dl_k. A linear regression through the origin of Poisson variables m_k against l_k was fitted and the generalized Pearson chi-square used as a measure of lack of fit (Collett 1991) to assess the evidence for nonrandomness.

We can also compare the observed number of segments per chromosome with a prediction from the model. To estimate the predicted number of observed segments the distribution $[y]_k$ is replaced by the distribution of the observed segments

$$[y_{\text{obs}}]_k = \text{Prob}((n|y) \geq 1) \times [y]_k$$

$$\text{with mean} = \int_0^{l_k} y[y_{\text{obs}}]_k \, dy.$$

Then S_k and *S* equivalents are calculated as before.

Gene mapping data from the chicken genetic linkage map: For chicken, the genes were mapped as part of the EC CHICKMAP project and the worldwide effort to map the chicken genome (Burt *et al.* 1995; Burt and Cheng 1998). The mapping information is recorded in the chicken genome database, Arkdb-chick (http://www.ri.bbsrc.ac.uk).

To estimate the genetic length of the chicken genome we take map lengths from recombination among *m* loci, using the Map Manager program (Manly 1993), corrected using the Kosambi mapping function (Kosambi 1944) and multiplied by $(m + 1)/(m - 1)$ to adjust for failure to sample telomeric regions (Morton 1991). The second correction assumes that loci are sampled randomly from a uniform distribution along the genetic map.

The locations of human and mouse genes were taken from the Genome Database (http://gdbwww.gdb.org/gdb/), UniGene (http://www.ncbi.nlm.nih.gov/), Online Mendelian Inheritance in Man (http://gdbwww.gdborg/

TABLE 1

Conserved genes and numbers of syntenic groups for chicken-human and chicken-mouse comparisons, together with the numbers of ranges defined by a conserved syntenic block with two or more genes

Observations	Chicken-human	Chicken-mouse
Conserved loci		
Random	132	119
Nonrandom	63	61
Density of random conserved loci (no./cM)	0.034	0.031
No. of chicken linkage groups with syntenic groups	28	26
Syntenic groups		
Single loci	41	67
Ranges	28	18
No. of ranges from a block of size		
$n = 2$	14	8
3	4	7
4	6	1
5	2	1
6	—	1
7	1	—
10	1	—

omim/docs/omimtop.html), and the Mouse Genome Database (http://www.informatics.jax.org/).

The comparative gene map for chicken, human, and mouse (http://www.ri.bbsrc.ac.uk) contains 214 orthologous loci, most of which are known genes or conserved anonymous cDNA sequences. We excluded members of multigene families or genes for which specific orthology could not be determined or for which homology was in doubt.

RESULTS

Data presented for comparative maps are based on chicken linkage groups and are labeled chicken-human (C-H) and chicken-mouse (C-M).

Data: Details of the observed numbers of conserved genes between chicken and human or mouse are given in Table 1. Gene density, for those considered found at random, was ~3/100 cM for both comparisons, having excluded one-third of the conserved loci that were considered nonrandom and therefore biased. The total estimated length of the linkage groups in the chicken map was 3836 cM. There were considerably more single loci than conserved syntenic groups with ranges, particularly so for the chicken-mouse comparison. Most of the ranges were derived from fewer than five genes. Ranges were observed on 19 (C-H) and 13 (C-M) linkage groups, and almost always as a single range per linkage group except for the four largest linkage groups (Figure 1). Smaller linkage groups were more likely to contain single loci than ranges for the chicken-mouse comparison. In all, 28 (C-H) and 26 (C-M) linkage groups were found to contain homology segments defined by a single gene ($n = 1$) or conserved syntenic groups with $n \geq 2$. The largest observed ranges from both comparative maps exceeded the median linkage group length, emphasizing the need for models allowing for chromosome size.

Tests of randomness, using the number of loci per linkage group, gave χ^2_{37} values of 33.4 (C-H) and 28.9 (C-M); neither provided evidence against randomness. The data and fitted lines representing the expected number of genes, assuming random scattering, are shown in Figure 2.

Model fitting and predictions: The results of fitting the various models to different data types are presented in Table 2 for the chicken-human comparison and in Table 3 for the chicken-mouse comparison. The behavior of the models was broadly similar for both comparative maps. Using either count data alone or combined data there was no evidence against the random cutting of chromosomes. In contrast, a model of nonrandom breakage was preferred when range data was considered on its own [$\chi^2_1 = 13.16$, $P < 0.001$ (C-H) and $\chi^2_1 = 7.12$, $P < 0.01$ (C-M) for a comparison of the three-parameter model with the two-parameter model for random chromosome breakage]. For both comparative maps the estimated numbers of segments from the nonrandom cuts model were less than the observed numbers of 69 (C-H) and 85 (C-M). Much of the information about the frequency of the short conserved segments is lost from the data when single loci are excluded. Estimates of the number of conserved segments for combined data are intermediate between those for range data and count data alone, but much closer to those obtained from counts. Confidence intervals derived from combined data were less than two-thirds the width of those from count data alone. This reflects both the extra information in ranges and the expectation that larger point estimates would give rise to wider confidence intervals.

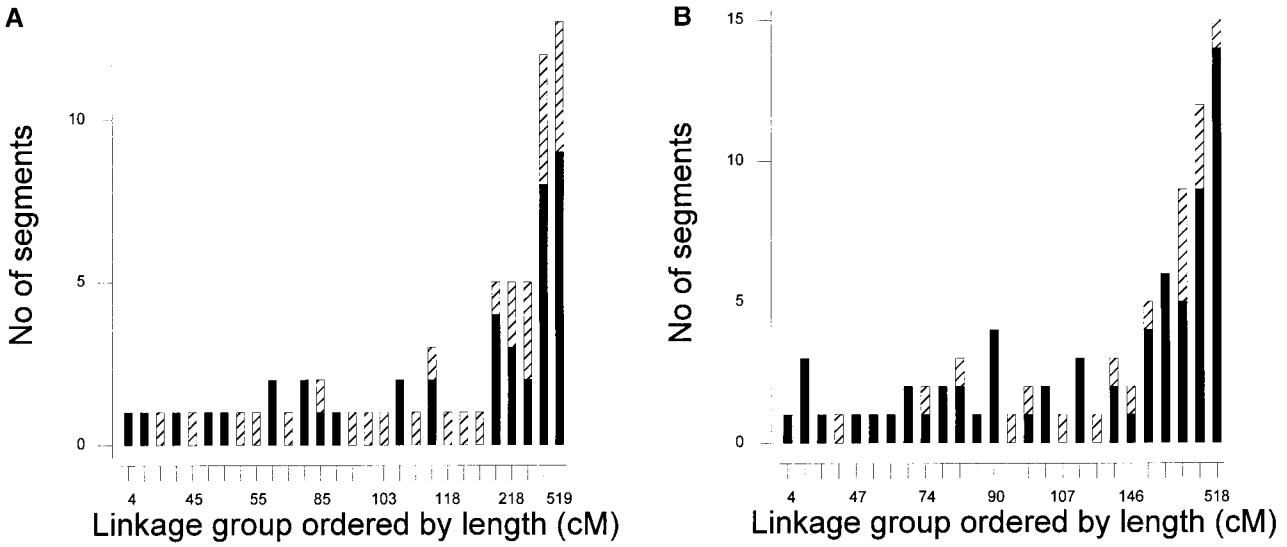

Figure 1.—Numbers of conserved syntenic blocks defined by one (solid) or more than one (hatched) gene observed in each chicken linkage group for (A) chicken-human (C-H) and (B) chicken-mouse (C-M) comparisons. Linkage groups are ordered by size.

The single-parameter random genome breakage model was favored (as the simplest model giving a comparable fit) in both comparative maps when using count or combined data, with the exception of the chicken-mouse combined data, where the two-parameter random chromosome breakage model was preferred. Both models give very similar estimates for the number of conserved segments and also have similar confidence intervals.

The observed numbers of segments per linkage group, plotted against linkage group length, are presented in Figure 3. Also included are predictions from the random chromosome breakage model of the number of underlying segments per linkage group and of the number of observed segments per linkage group with a 95% confidence interval. The chicken-mouse prediction for the number of observed segments shows good agreement with the data. Even so, there are still some (nonzero) observed numbers outside the confidence range for observed segments. This is inevitable when the model predicts a single segment for very small linkage groups. For the same reason the predicted curve for the number of conserved segments also lies below some of the observed numbers of segments for short linkage

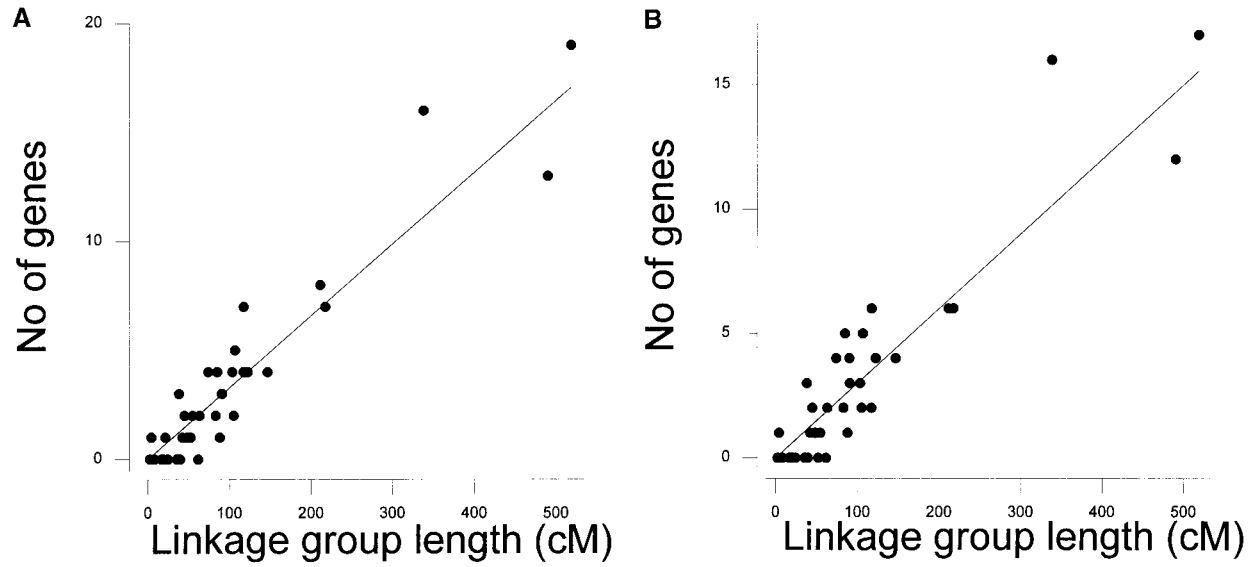

Figure 2.—Numbers of conserved loci per chicken linkage group *vs.* linkage group length, together with fitted lines corresponding to uniform gene density, for the (A) chicken-human (C-H) and (B) chicken-mouse (C-M) tests for random scattering of genes.

TABLE 2

Fits of various chromosome breakage models, together with maximum-likelihood estimates of parameters and confidence intervals, using count, range, and combined data, for chicken-human conserved segments

Data type	Chromosome breakage model	Maximum log-likelihood	$\hat{\alpha}^a$	β^a	$\hat{\gamma}$	$\hat{\delta}^a$	$\hat{S}$	95% C.I. for $\hat{S}^b$
Counts	Nonrandom	$L_1 = -90.50$	0.75	0	0.0020	1.38	111	—
Counts	Random chromosome	$L_1 = -90.72$	1	0	0.0017	1.45	110	(78, 156)
Counts	Random genome[c]	$L_1 = -92.62$	1	0	0.0197	1	114	(88, 150)
Ranges	Nonrandom[c]	$L_2 = -171.70$	18.82	0	0.00023	2.08	55	—
Ranges	Random chromosome	$L_2 = -178.28$	1	0	0.00050	1.49	64	—
Ranges	Random genome	$L_2 = -179.70$	1	0	0.00657	1	63	—
Combined	Nonrandom	$L_3 = -213.44$	0.64	0	0.0029	1.25	102	—
Combined	Random chromosome	$L_3 = -214.34$	1	0	0.0023	1.36	95	(75, 122)
Combined	Random genome[c]	$L_3 = -215.94$	1	0	0.0156	1	98	(81, 120)

For explanation of symbols see methods. C.I., confidence interval.
[a] Parameter values of 0 or 1 are fixed.
[b] For random breakage models only, if supported.
[c] Suggested model for each data type from comparison of log-likelihoods.

groups. In the chicken-human comparison this also occurs for the longest linkage groups.

An illustration of the flexibility of the β distribution models to represent a wide range of segment length distributions is presented in Figure 4 using fitted distributions corresponding to the estimated parameters from the chicken-mouse comparison. The upper diagram shows changes in segment length distributions with chromosome length from the random chromosome breakage model applied to the combined data. Four chromosome lengths have been chosen for illustration. The distribution for the shortest chromosome of 20 cM has the most probable segment length equal to the chromosome. At a length of 60 cM the distribution is almost uniform, which would be appropriate for a single random cut point. At 90 cM the distribution becomes triangular, corresponding to two cut points. Longer chromosomes show an exponential-like segment length distribution shifted progressively further to the left. This is shown for a chromosome length of 150 cM, corresponding to 5 segments, for which the probability of a segment exceeding half of the length of the chromosome is 0.06. This probability halves for each additional segment on a chromosome.

The lower diagram in Figure 4 is of the nonrandom breakage model fitted to the range data alone. For the shortest chromosome the most probable segment length is equal to that of the chromosome, as in the random breakage model. As chromosome lengths increase, however, the segment length distributions have progres-

TABLE 3

Fits of various chromosome breakage models, together with maximum-likelihood estimates of parameters and confidence intervals, using count, range, and combined data, for chicken-mouse conserved segments

Data type	Chromosome breakage model	Maximum log-likelihood	$\hat{\alpha}^a$	$\hat{\beta}^a$	$\hat{\gamma}$	$\hat{\delta}^a$	$\hat{S}$	95% C.I. for $\hat{S}^b$
Counts	Nonrandom	$L_1 = -69.74$	0.71	0	0.012	1.16	193	—
Counts	Random chromosome	$L_1 = -70.09$	1	0	0.014	1.21	190	(126, 305)
Counts	Random genome[c]	$L_1 = -70.44$	1	0	0.041	1	195	(141, 282)
Ranges	Nonrandom[c]	$L_2 = -100.75$	7.84	0	0.00002	2.49	74	—
Ranges	Random chromosome	$L_2 = -104.31$	1	0	0.00006	1.94	76	—
Ranges	Random genome	$L_2 = -106.87$	1	0	0.0081	1	69	—
Combined	Nonrandom	$L_3 = -140.56$	0.59	0	0.0028	1.38	170	—
Combined	Random chromosome[c]	$L_3 = -142.38$	1	0	0.0021	1.51	155	(113, 219)
Combined	Random genome	$L_3 = -145.00$	1	0	0.0296	1	152	(119, 194)

For explanation of symbols see methods. C.I., confidence interval.
[a] Parameter values of 0 or 1 are fixed.
[b] For random breakage models only, if supported.
[c] Suggested model for each data type from comparison of log-likelihoods.

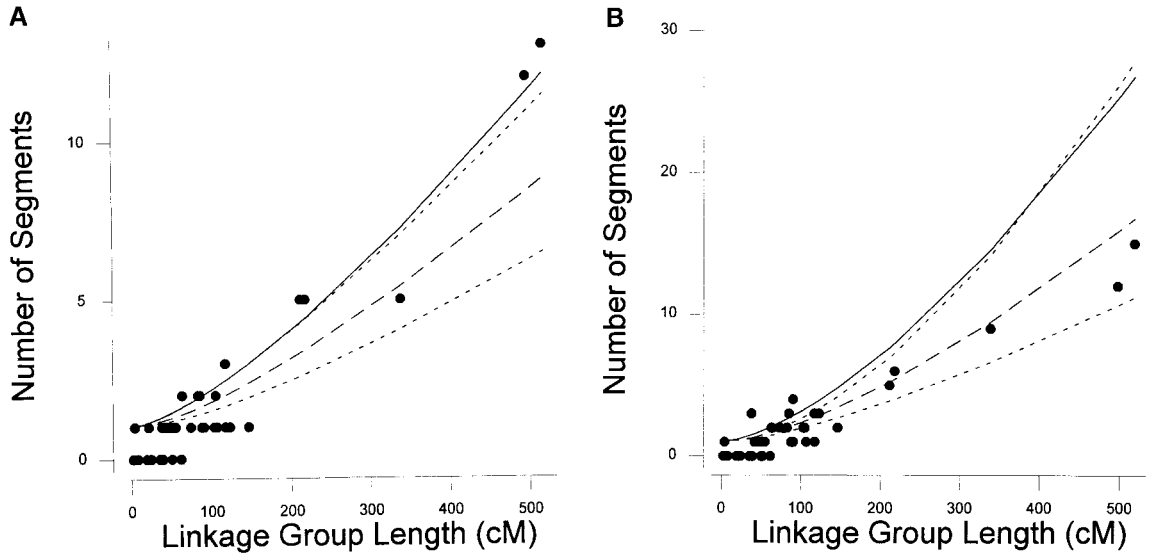

Figure 3.—Numbers of observed conserved segments per chicken linkage group *vs.* linkage group length for the (A) chicken-human (C-H) and (B) the chicken-mouse (C-M) comparisons. Predicted numbers of conserved segments from the random chromosome breakage model using combined data (solid line), and the corresponding prediction of observed numbers of segments (dashed line), together with its 95% confidence limits (dotted lines).

sively smaller means relative to the length of the chromosome and are unimodal. There is no evidence in the range data, and therefore no reflection in the fitted distributions, of a preponderance of very small segment lengths.

DISCUSSION

The small size of some chicken chromosomes and the relatively large size of some conserved syntenic blocks have driven the construction of a chromosome-based model for conserved segments. But, as the model is a generalization of that of Nadeau and Taylor, there is no reason why the approach should not be used more widely, particularly with its emphasis on the testing of model and data assumptions. As illustrated in results, the β distribution has provided a very flexible and intuitively appealing range of distributional shapes for the unobservable segments. Particularly important are special cases corresponding to random breakage models, one of which is already prevalent in the literature (see Nadeau and Sankoff 1998 for a review). The likelihood approach presented here allows an explicit test of the plausibility of these random-breakage models, as well as providing a framework for deriving the confidence intervals that are an essential accompaniment to estimates. A particularly striking consequence of using such a flexible model is the need to use all available data to draw reliable conclusions. Discarding segments defined by single loci (homology segments) results in gross underestimation of genomic rearrangement. These issues are discussed in more detail below.

There are two untested assumptions made in the Nadeau and Taylor (1984) model, which have also been made here, without extensive comment: that genetic and physical distances are (approximately) proportional and that there are no insertions within a syntenic block, although inversions are permitted and fairly common in large conserved segments. This will lead to underestimation of genomic rearrangement. The algorithm of Sankoff *et al.* (1997a) could be used to identify probable inversions in the data, rather than those caused by incorrect gene ordering, prior to model estimation.

The crucial assumption of genes spread at random over the genome has been tested, but at the simplest level, to assess constant density over chromosomes. There is some evidence that recombination rates in the chicken microchromosomes (the smallest 33 autosomes) are some 2.5 times those in macrochromosomes (the largest 5 autosomes; Rodionov 1996), and that gene densitites in microchromosomes are double those in macrochromosomes (Smith *et al.* 1999). These two effects cancel out in the test for a random scattering of genes. Performing the same test on the human-mouse comparative map using >1600 unselected genes gave a χ^2_{18} value of 142, indicating a range of gene densities on different chromosomes well in excess of expectation under randomness. For this comparative map, the removal of nonrandomly selected genes represents a formidable task. Of course, assuming random scattering of the genes will only be an approximation, but a very useful one that is likely to become increasingly untenable, and practically impossible to remedy, as maps become more detailed. The overall consequences of nonrandomness of genes are not easily predicted. If genes

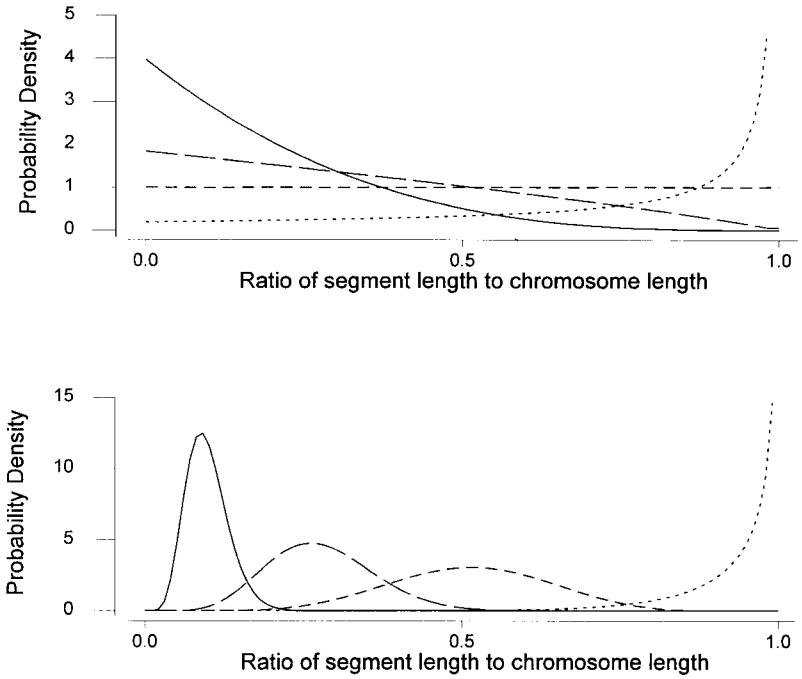

Figure 4.—Changes in the distribution of the ratio of segment length to chromosome length for the chicken-mouse comparison. Chromosomes are scaled to have length = 1 for display on a common axis. The upper diagram is from the fit of the random chromosome breakage model applied to the combined data, showing chromosomes of length 20 (dots), 60 (dashes), 90 (long dashes), and 150 cM. The lower diagram is from the fit of the nonrandom breakage model applied to range data alone, for chromosomes of length 60, 180, 270, and 450 cM, with line style order as above.

are too clustered (as a result of chromosome walking or proximity of gene families), then corresponding segment lengths will be overestimated, resulting in a downward bias for the estimated number of conserved segments. If genes are too evenly spread (perhaps by map cross-referencing), then an excess of short homology segments may be observed, leading to an upwardly biased estimate. A preliminary model allowing for nonrandom gene distributions has recently been proposed by Sankoff *et al.* (1997b), but further development is needed. For comparative mapping purposes, linking species with nascent maps to the detailed maps of humans or mice will be highly beneficial, but these are the very maps where nonrandomness of genes will be unavoidable. With careful examination, randomness of gene discovery may be a plausible approximation in the newly mapped species, but the background densities of orthologous genes on mouse or human maps may well vary. If this variation leads, as a first approximation, to groups of chromosomes of similar densities, our model is easily modified to reflect this.

The random chromosome and genome breakage models presented here are obtained as special cases of an empirical nonrandom breakage model. This allows likelihood-ratio tests for independent components of the model, an approach that is preferable to using goodness-of-fit tests for the whole model and then, if satisfactory, declaring that all of the model components are validated. Conclusions about the pattern of chromosome breakage are strongly influenced by the choice of which data measurements to analyze. It may be that when using only the ranges, which contain indirect metric evidence about segment lengths, we are detecting genuine nonrandomness. Or, perhaps, short ranges are underrepresented in the relatively small number of ranges in our sample covering the whole genome, often resulting in just a single range being present on a linkage group. When using all the data the tests within the model do not provide evidence against "chromosomes" being cut at random for the two comparative maps presented here, although the evidence is not unanimous about whether the randomness is on a chromosome or a genome basis. However, both models give similar estimates of conserved numbers of segments. Furthermore, if the model is modified so that observed ranges and linkage group lengths corresponding to microchromosomes are reduced by a factor of 2.5 (the minimum shrinkage factor corresponding to the almost linear part of the Kosambi mapping function), then the random genome breakage model is preferred for both comparative maps, and the estimates of conserved segment numbers change little. Other evidence for random genome breakage models is presented in Nadeau and Sankoff (1998). One arbitrary feature of the chromosome model presented here is the smooth relationship chosen to change the distribution of segment length with chromosome length. There is no expectation that the number of conserved segments on a chromosome increases monotonically with chromosome length, although when chromosome lengths differ widely an increasing trend is likely. With random genome breakage the trend should be linear, becoming less variable with an increase in the number of generations and rearrangements between the two species being compared. This may be a factor in the superior agreement of the observed segment numbers and their prediction in the chicken-mouse

comparison. The chicken-human comparison is dominated by linkage groups with only a single observed segment (Figure 1), and this suggests that other functions might be useful in relating the number of conserved segments to chromosome length for some contexts. For example, the model may be easily modified to fit different breakage rates among microchromosomes and among macrochrosomosomes in chickens, if considered biologically plausible.

The estimates of the number of conserved segments change considerably depending on which measurements are chosen as data, in contrast to the relative stability of the estimates over the different chromosome breakage models. The flexibility of these breakage models in describing segment length distributions means that the model will be more sensitive to the data than might be the case with, say, a single-parameter exponential distribution. This places considerable emphasis on the quality of the mapped data and the examination of the assumptions used to describe gene distributions.

As an example of the dramatic effect of very different model assumptions on conclusions, we have also used the model of Sankoff *et al.* (1997b) to estimate conserved segment number. This elegant model, derived from the probability theory of runs, treats the genome as continuous and relies on both the process of gene identification and chromosome breakage occurring at random. It excludes considerations of gene order and distance both between and within conserved segments, and consequently ignores the impossibility of some configurations of observed ranges and chromosome boundaries. Further work is required to assess the limitations of the model assumptions. Estimates of the conserved number of segments from this model are considerably larger than those derived from our models; C-H = 142 segments and C-M = 293 segments.

Using our models, the chicken-mouse comparison gave an estimate of the number of conserved segments of almost 50% more than the chicken-human comparison for the combined data, with a confidence interval twice as wide. The intervals barely overlapped, suggesting a difference in conserved number of segments with the chicken between these two species. Further evidence of a difference comes from an examination of the ranges that were found in both comparisons. There are 16 ranges in common, of which 8 were of equal length. The remaining 8 have a measured range that is shorter for the chicken-mouse comparison. The two comparative maps presented here have many common genes in their data: genes that are first mapped in the chicken and then located in both of the more extensive mouse and human maps. This precludes the simplest approach to testing for human and mouse differences in conserved segment numbers with chicken by pooling data and assuming it to be independent, because there may be a positive correlation between the estimates of conserved segment number caused by the sampling scheme above. Finding a satisfactory representation of this correlation will be important in future work in evolutionary modeling, because in the long term we will wish to assess differences in the number of conserved segments for multiple comparative maps and to use these maps to give a new perspective on phylogenetic trees.

We thank Liz Archibald for her excellent typing, and Michael Romanov for Russian translation. We also thank the Ministry of Agriculture, Fisheries and Food (MAFF), the Biotechnology and Biological Sciences Research Council (BBSRC) and the Commission of the European Communities for supporting this work.

LITERATURE CITED

Andersson, L., M. Ashburner, S. Audun, W. Barendse, J. Bitgood *et al.*, 1996 Comparative genome organization of vertebrates: the first international workshop on comparative genome organization. Mamm. Genome **7:** 717–734.

Bengtsson, B. O., K. Klinga Levan and G. Levan, 1993 Measuring genome reorganization from synteny data. Cytogenet. Cell Genet. **64:** 198–200.

Burt, D. W., and H. H. Cheng, 1998 Chicken gene map. ILAR J. **39:** 185–192.

Burt, D. W., N. Bumstead, J. J. Bitgood, F. A. Ponce de Leon and L. B. Crittenden, 1995 Chicken genome mapping: a new era in avian genetics. Trends Genet. **11:** 190–194.

Burt, D. W., C. Bruley, I. Dunn, C. T. Jones, A. S. Law *et al.*, 1999 Dynamics of chromosome evolution in birds and mammals. Nature (in press).

Carver, E. A., and L. Stubbs, 1997 Zooming in on the human-mouse comparative map: genome conservation re-examined on a high-resolution scale. Genome Res. **7:** 1123–1137.

Collett, D., 1991 *Modelling Binary Data*. Chapman and Hall, London.

Erlich, J., D. Sankoff and J. H. Nadeau, 1997 Synteny conservation and chromosome rearrangements during mammalian evolution. Genetics **147:** 289–296.

Hannenhalli, S., 1995 Polynomial time algorithm for computing translocation distance between genomes, pp. 162–176 in *Combinational Pattern Matching, 6th Annual Symposium, Lecture Notes in Computer Science*, edited by Z. Galil and E. Ukkonen. Springer-Verlag, New York.

Hannenhalli, S., and P. A. Pevzner, 1995 Transforming men into mice (polynomial algorithm for genomic distance problem). Proceedings of the 36th Ann. Symposium Found. Comp. Sci., IEEE Computer Society Press, pp. 581–592.

Jones, C. T., D. R. Morrice, I. R. Paton and D. W. Burt, 1997 Homologues of genes on human chromosome 15q21-q26 and a chicken microchromosome show conserved synteny and gene order. Mamm. Genome **8:** 436–440.

Kosambi, D. D., 1944 The estimation of map distance from recombination values. Ann. Eugen. **12:** 172–175.

Manly, K. F., 1993 A Macintosh program for storage and analysis of experimental genetic mapping data. Mamm. Genome **4:** 303–313.

Morizot, D. C., 1983 Tracing linkage groups from fishes to mammals. J. Hered. **74:** 413–416.

Morton, N. E., 1991 Parameters of the human genome. Proc. Natl. Acad. Sci. USA **88:** 7474–7476.

Nadeau, J. H., and D. Sankoff, 1998 Counting on comparative maps. Trends Genet. **14:** 495–501.

Nadeau, J. H., and B. A. Taylor, 1984 Lengths of chromosomal segments conserved since divergence of man and mouse. Proc. Natl. Acad. Sci. USA **81:** 814–818.

O'Brien, S. J., J. E. Womack, L. A. Lyons, K. J. Moore, N. A. Jenkins *et al.*, 1993 Anchored reference loci for comparative genome mapping in mammals. Nat. Genet. **3:** 103–112.

O'Brien, S. J., J. Wienberg and L. A. Lyons, 1997 Comparative genomics: lessons from cats. Trends Genet. **13:** 393–399.

Pitel, F., V. Fillon, C. Heimel, N. Le Fur, C. El Khadir-Mounier *et al.*, 1998 Mapping of FASN and ACACA on two chicken mi-

crochromosomes disrupts the human 17q syntenic group well conserved in mammals. Mamm. Genome **9:** 297–300.

Plackett, R. L., 1971 *An Introduction to the Theory of Statistics.* Oliver and Boyd, Edinburgh.

Postlethwait, J. H., Y.-L. Yan, M. A. Gates, S. Horne, A. Amores *et al.*, 1998 Vertebrate genome evolution and the zebrafish gene map. Nat. Genet. **18:** 345–349.

Rodionov, A. V., 1996 Micro versus macro: a review of structure and functions of avian micro- and macrochromosomes. Genetika **32:** 597–608.

Sankoff, D., 1993 Models and analyses of genomic evolution, pp. 177–183 in *Supercomputing and Complex Genome Analysis, Proceedings of the Second International Conference on Bioinformatics,* edited by H. A. Lim, J. W. Fickett and C. R. Cantor. World Scientific Publishing Co. Pte. Ltd., Singapore.

Sankoff, D., and J. H. Nadeau, 1996 Conserved synteny as a measure of genomic distance. Discrete Appl. Math. **71:** 247–257.

Sankoff, D., V. Ferretti and J. H. Nadeau, 1997a Conserved segment identification. J. Comput. Biol. **4:** 559–565.

Sankoff, D., M.-N. Parent, I. Marchand and V. Ferretti, 1997b On the Nadeau-Taylor theory of conserved chromosome segments, pp. 262–274 in *Combinational Pattern Matching, 8th Annual Symposium, Lecture Notes in Computer Science*, edited by A. Apostolico and J. Hein. Springer-Verlag, New York.

Smith, E. J., and H. Cheng, 1998 Mapping chicken genes using preferential amplification of specific alleles. Microbial and comparative genomics. Genomics **30:** 13–20.

Smith, J., C. K. Bruley, I. R. Paton, I. Dunn, C. T. Jones *et al.*, 1999 Differences in gene density in the chicken macrochromosomes and microchromosomes. Anim. Genet. (in press).

Womack, J. E., and S. Kata, 1995 Bovine genome mapping: evolutionary inference and the power of comparative genomics. Curr. Opin. Genet. Dev. **5:** 725–733.

Zakharov, I. A., V. S. Nikiforov and E. V. Stepanyuk, 1995 Interval estimates of the combinatorial measures of similarity for orders of homologous genes. Genetika **31:** 1163–1167.

Communicating editor: G. A. Churchill